DOCUMENT RESUME

ED 206 658 TM 810 550

AUTHOR Christie, Samuel G.; Conniff,, William A.
TITLE A Case in Support of Using Locally Developed Non-Normed Tests for Title I Program Evaluation.
PUB DATE 15 Apr 81
NOTE 20p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Elementary Secondary Education; Error of Measurement; *Norm Referenced Tests; *Program Evaluation; *School District Autonomy; *Test Construction; Test Reliability
IDENTIFIERS Contrasting Groups Method; Equipercentile Equating; Stockton Unified School District CA; *Title I Evaluation and Reporting System

ABSTRACT

Stockton Unified School District's successful strategy of using a locally developed, non-normed achievement test to implement the norm referenced model of the Title I Evaluation and Reporting System (Model A2) is described. Documented are the procedures involved in the development of a curriculum guide and test items, and the administration of field tests and subsequent revisions. The equipercentile method used for equating the local test with national tests to show its reliability, and the contrasting groups method, conducted to set the passing score, are also outlined and given justification. The locally developed test and the national norm-referenced test were administered to the total student population at each grade level over a period of two weeks. The resultant statistics, reliability coefficient and concurrent validity coefficients are presented in tables. (AEF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED206665

NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Case in Support of Using Locally Developed Non-Normed Tests for Title I Program Evaluation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. G. Christie
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Samuel G. Christie
and
William A. Conniff

Stockton Unified School District

TM 810 550

A Paper Presented at the
American Educational Research Association Annual Meeting
Los Angeles, California
April 15, 1981

The purpose of the project described in this paper was: 1) to meet the requirement for evaluating a Title I program, 2) to do this with a minimum amount of testing time, and 3) to produce data which teachers would perceive as being relevant to their instructional program.

This paper reports on the approach taken by one school district to meet these objectives. The strategy in this district was to use a locally developed test to implement the norm-referenced model (Model A2) of the Title I Evaluation and Reporting System. As so often happens in a school district, systematic planning is a seldom realized hope, happening only when the circumstances are right. In the case in point, the development of new evaluation models by the Federal government happened to be completed at about the same time that field testing on newly developed local tests were being completed.

The adoption of local tests came at the end of three years of developmental work. In 1976, the Board of Education of Stockton Unified School District in California directed the Superintendent to have a Language Arts Scope and Sequence (LASS) developed and to set grade level expectations in reading and written language. The following year development of a test to measure these expectations was begun. At the same time, a task force of teachers led by the district curriculum coordinator began to develop a Math Scope and Sequence (MSS). The following year development of the math test began.

This move to develop local tests was taken for a number of reasons. The teachers had long complained that the standardized test administered in the district was not testing all the things they taught during a particular

year. That is, the curriculum and the tests were out of sequence. The Board wanted to have parents better informed about their children's achievement. It was felt that a locally developed test to measure achievement, one that was congruent with the curriculum in use, would not only help teachers to communicate the level of achievement to parents, it could also help in the improvement of instruction.

Development of the Curriculum Guide

Prior to test development, grade level curricula had to be developed. This was done in what has become rather standard practice in school districts. That is, a central office person contacted the schools informing them that a committee would be formed and announcing the purpose of the committee.

Principals in each school were asked to pick their most qualified teacher to work on this committee.

The charge given to this committee was to identify skills and knowledge that would be expected of students. The question that guided this work was: As students enter each grade, kindergarten through sixth, what are the skills and knowledge that they must possess if they are going to be able to progress in the new grade.

The committee examined textbooks and other curriculum materials that were being used at that time as well as listening to teachers who told them what they thought should be included in the guide.

Initial lists of expectations for the grade levels were ciruculated for additional input. Through this process, a continuum for language arts and for math was developed.

Development of the Local Test

Original test items were written by the coordinator of curriculum and teachers on a test writing committee. The evaluation office had only given general guidelines about writing test items before this work was begun.

The items written by this committee were revised by the evaluation office, with few corrections made. The corrections that were made dealt primarily with format. This version was then field tested.

After the first field test, the evaluators became more involved. An item analysis was conducted utilizing a computer program that had been written sometime before this test development was contemplated. This program correlated the top and bottom 27% of the responses to the total test score to determine item discrimination. On the basis of this analysis, revisions were made.

In the revisions, a few items were rewritten. The number of choices offered as alternatives varied with some questions requiring three choices, other four. The original item writers seemed to run out of ideas for alternate choices as most questions had three choices instead of four. In the revision, the questions were made consistent with four alternatives written for each question. The most difficult items to develop were those measuring reading comprehension. Not only were the items more difficult to construct, there were more of them, as reading comprehension was emphasized at all the grade levels.

After the first revisions were completed, the tests were administered again and reanalyzed. This time only minor revisions were made, as the test statistics were quite good. (A sample of this analysis is shown in the appendix).

All this developmental work was going on with no thought given to making these tests a part of the Titel I Evaluation and Reporting System.

The district had committed itself to implementing a locally developed achievement testing program prior to learning of the ESEA Title I evaluation models. Originally, the plan had been to test fall-spring with the local test and spring-spring with the nationally normed test.

We had considered spring to spring testing with both the local and the national test but decided against it in favor of a fall-spring schedule for the local test because it provided us with better data for planning and evaluation. We realized that this decision to have these test administrations yearly would not be popular with teachers, therefore, when we learned of the possibility of dropping the Metropolitan

Test we were interested. We felt we would have better data with our local test, also we could better concentrate everyone's attention on grade level expectations as the school board had wanted and we could cut back to only two test administrations per year.

The norm-referenced evaluation model (A2) permits the use of a non-normed (local) achievement test (NNT), however, it requires the use of a nationally norm referenced test (NRT) periodically to equate the local test scores to national percentiles. These percentiles can then be converted to norm curve equivilent (NCE) scores. NCE scores were adopted for use by the Office of Education because these scores will satisfy four criteria which are important in terms of measuring the impact of Title I projects. NCE's may be compared across achievement levels, they are not difficult to interpret, they are equal interval scores, and they accurately measure both small and large gains.

NCE scores are obtained by converting national percentile scores to NCE scores using a standard conversion table. In order to obtain national percentiles when an NNT is used, the NNT is statistically equated to an NRT. The process of equating the NNT to an NRT yields national percentiles which are statistically equivalent to the NNT raw scores. This enables one to develop a table whereby raw scores on the local test can easily be read out in terms of a national percentile.

Equating the National and Local Tests

There are four common approaches to test equating:

1) linear transformation

2) curvilinear transformation (equipercentile)

3) common or anchor test equating

4) latent trait linking

The equipercentile method of equating was selected over the three other approaches for the following reasons: The linear transformation approach produces approximately the same results as the equipercentile approach, provided that the NNT

and NRT distributions are nearly identical. Examination of the two distributions indicated that these two distributions were similar but not sufficiently so to use the linear method. This is supported by the magnitude of the Pearson Product - Moment correlation coefficients which ranged from .82 to .87 in grades 2 - 6 reading and .76 to .83 in math. Anchor test equating was not chosen because it has more chances for increased error. Latent trait procedures were not used to develop the NNT and could not, therefore, be used to equate the NNT to the NRT.

The principle underlying the equipercentile method of equating is that raw scores on the test being equated are considered equivalent if they have the same local percentile based on a concurrent administration of the tests to the same sample. Tests equated on this basis are statistically equivalent only. In order to use the NNT for program evaluation, it was necessary to establish content and concurrent validity of the NNT in order that the equated national percentiles could be interpreted similar to the NRT national percentiles.

Design

The NNT and the NRT were administered at approximately the same time (within a two-week period) to the total population of students at each grade level in the spring of 1980. A counter-balanced design was employed

where one-half of the classrooms at each grade level was randomly selected to take the NNT followed by the NRT and the remaining one-half was scheduled to take the NRT first followed by the NNT. The NNT and the NRT were given at posttest time; i.e., the time at which the NRT was normally scheduled to be given.

An NNT-NRT conversion table was developed based on the 1980 equating study and will be used to convert both fall and spring NNT scores to national percenitles. The equating study will be repeated every three years. This will serve to reduce error resulting from demographic changes in the student population as well as to provide national percentiles which are free from any error due to equating.

Implementation

The Title I Technical Assistance Center (TAC) for region 9 and the California State Department of Education (CSDE) oversaw SUSD's equating study and stipulated that the following criteria be met.

1) The NNT must exhibit good item and test characteristics and the content of the NNT must be similar to the NRT.
2) The NNT must have concurrent validity as determined by a high correlation between the NNT and the NRT.
3) The NNT must be a reliable instrument as demonstrated by suitable reliability coefficients.
4) SUSD must construct NNT to NRT equating tables based on the proper procedure.

The first criteria was met based on a review by the TAC and SDE personnel of 1) a copy of each level of the NNT; 2) a comprehensive item analysis for each level of the NNT based on the Spring 1980 SUSD districtwide

administration (sample item analysis appears in the appendix); 3) the test content outline for the NNT which lists each skill and the corresponding test items; and 4) KR20 reliability coefficiencts,[1] which are given in Table 2.

The second criteria was met based on the Pearson Product-Moment correlations given in Table 1. As would be expected, the NNT-NRT correlation coefficients were lowest at kindergarten where the KR20 coefficients were also lowest (Kindergarten and first grade scores are not reported for Title I evaluation).

The completed NNT-NRT conversion tables were submitted to the TAC for review and were approved. The conversion table for grade four is in Table 3.

Sources of Error

In addition to the standard error of measurement which is largely a question of the reliability of a test, equated tests have an additional source of error known as the standard error due to equating. The standard error due to equating is the standard deviation of the equated scores on the scale of the NRT corresponding to a fixed score on the NNT, where the equated scores are derived from a conversion line which results from the distribution on the NNT and the NRT based on the same group. A conversion line is formed by connecting points which represent the same local percentile on the two distributions. The standard error due to equating is inversely related to the size of the equating sample and the correlation between the NNT and the NRT and is directly related to the distance of the score point from the mean.

Each of the first two factors was high. The entire population at each grade level was used for the equating study and ranged from 1,558 at grade six to 1,638 at grade four. The Pearson Product-Moment correlations were

1. For an explanation of the choice of the KR20 coefficeient, see (Conniff, W., 1980) in the References.

TABLE 2

KR20 Coefficients for the NNT
for grades K - 6, Reading and Mathematics

NNT

	Reading	Mathematics
K	.879768	.834302
1	.919524	.907850
2	.954776	.923512
3	.947679	.903485
4	.950719	.919527
5	.948471	.902579
6	.949417	.931241

TABLE 1

Pearson Product Correlations for the NNT
for grades K - 6, Reading and Mathematics

NNT

	Reading	Mathematics
K	.74282	.69562
1	.83767	.72058
2	.86837	.76396
3	85820	.80883
4	.84740	.81712
5	.82075	.82696
6	.82810	.81255

TABLE 3

Portion of NNT - NRT Equating Table
grade 4, Reading

NNT		NRT	
Raw Score	Local Percentile	Local Percentile	Raw Score
89	99.0+	99.0+	92-94
88	99.0+	99.0+	91
87	99.0+	99.0+	90
86	98.8	98.9	89
85	98.1	98.0	88
84	97.2	96.9	87
83	96.1	95.8	86
82	94.8	94.4	85
81	93.2	93.0	84
80	91.1	91.4	83
79	88.9	89.6	82
78	86.7	86.1	80
77	84.5	83.8	79
76	82.6	81.5	78
75	81.0	81.5	78
74	79.4	79.3	77
73	76.9	77.3	76
72	74.4	75.3	75
71	72.1	71.5	73
70	69.6	69.6	72
69	67.6	67.6	71
68	65.6	65.5	70
67	63.7	63.1	69
66	62.2	63.1	69
65	60.6	60.9	68
64	58.7	58.9	67
63	56.7	57.1	66

high, most being above .80. The third factor varies from score to score related to the regression toward the mean phenomenon. This operates such that the farther a score is from the mean the greater the error due to equating. Hence, different equipercentile scores have different amounts of error due to equating.

Table 4 gives the standard error due to equating corresponding to every fifth raw score point for each grade level of the reading NNT and the math NNT. The standard error due to equating was less than one point in each case. The formula for the standard error due to equating is:

$$SE_y = \sqrt{Sy^2 \,(1- r_{xy}) \;\frac{Zx^2 \,(1+ r_{xy}) + 2}{N_t}}$$

where:

Sy^2 = Variance of test Y

r_{xy} = Correlation between tests X and Y

Zx^2 = Square of the standard Z score for score X

N_t = Size of the sample used in the equating

In spite of large sample sizes and moderately high correlation coefficients, the amount of error in the equipercentile scores was often quite large. One can obtain some idea of the relative magnitude of error in the equipercentiles from Table 5. Table 5 contains the mean equipercentile score and the NRT mean percentile score for each school for grade 6, reading. These scores are from the concurrent administration of the NNT and the NRT in April, 1980. The mean NRT percentile for Kennedy school was fourteen points higher than Victory school (42 vs. 28). However, Kennedy's mean equipercentile score was two points lower than Victory's (32 vs. 34). This example represents the largest discrepancy found in the data for grades K-6 for reading and math. Discrepancies as large as six points were common.

Standard Error Due to Equating for each grade level
of the Reading and Math NNT

READING

SCORE	K	1	2	3	4	5	6
5	.37	.27	.50	.46	.72	.70	.64
10	.29	.22	.45	.41	.66	.64	.58
15	.22	.17	.39	.35	.60	.57	.51
20	.16	.13	.34	.30	.54	.51	.45
25	.12	.13	.29	.25	.48	.45	.40
30	.14	.16	.25	.21	.43	.39	.34
35	.19	.20	.21	.18	.38	.34	.30
40		.26	.19	.17	.33	.30	.27
45			.19	.18	.29	.27	.25
50			.21	.21	.26	.26	.25
55			.24	.25	.25	.27	.27
60			.28	.31	.25	.30	.31
65					.27	.34	.35
70					.30	.39	.41
75					.34		
80					.39		

MATH

SCORE	K	1	2	3	4	5	6
5	.25	.33	.53	.31	.31	.27	.36
10	.16	.25	.45	.24	.25	.21	.30
15	.10	.18	.38	.18	.19	.16	.24
20	.09	.13	.31	.14	.15	.13	.20
25	.15	.12	.25	.12	.13	.14	.18
30		.17	.20	.14	.15	.17	.19
35			.17	.18	.19	.22	.23
40			.18	.24		.28	.28
45			.22	.31			.33
50			.28	.37			.40

TABLE 5

Mean NRT Percentile Ranks and Mean NNT Equipercentile Ranks for grade 6, Reading

School	NRT	NNT
August	28	32
Cleveland	40	36
El Dorado	42	36
Fillmore	34	32
Garfield	16	16
Grunsky	26	23
Hazelton	32	28
Kennedy	42	32
Kohl	40	42
Madison	28	26
Montezuma	26	26
Roosevelt	32	23
Skills	66	72
Taylor	32	22
Valenzuela	26	22
Van Buren	18	22
Victory	28	34

One final step remained in the development of the local tests. We needed to determine how difficult they were in comparison to an extended standard.

Setting Grade Level (Passing) Scores

Since one of the purposes in developing the local tests was to evaluate grade level achievement, it was necessary to set a criterion level for each grade, i.e., to establish passing scores for each test. This was done during the time the equating study was carried out. The contrasting groups method was used to set the passing scores. Prior to the first administration of the adopted test, a sample of teachers at each grade level was asked to participate in this study. Teachers selected were those who were perceived to be most aware of the achievement of their students.

Each teacher was asked to categorize his/her students in two lists, those on grade level and those below grade level. The teacher was instructed to consider a student as being on grade level if the student was capable of making consistent progress working with the materials at his/her assigned grade level.

The list of names for each group of students were then returned to the research department. Two weeks later, the tests were administered and the test results of these two different groups of students were plotted. The point of overlap of these two distributions served as an empirical referent to suggest a grade level score on each test. This overlap would occur where some of those nominated as being below grade level had scored above some of the students who had been nominated as being on grade level.

The teacher's judgment proved to be quite consistent with the test results as there were only a few students in the parts of the distributions that overlapped (less than 10%). Comparisons of the grade level (passing)

score on the local test to the Metropolitan Test showed that grade level scores corresponded quite closely to the 50th percentile (see Table 6). For example, first grade students who scored 60 on reading (the passing score) scored at the 50th percentile. The passing score was within four percenitles of the 50th in all but one of the fourteen combinations of grades and subjects who achieved a passing score (76) scored at the 44th percentile.

TABLE 6

Comparison of Grade Level Scores on LASS and MSS Tests to Percentile Scores on the Metropolitan Achievement Test

Grade	Grade Level Score on LASS (percent correct)	Metro Equivalent Percentile	Grade Level Score on MSS (percent correct)	Metro Equivalent Percentile
K	77	47	83	—
1	60	50	83	50
2	72	48	74	50
3	75	46	76	44
4	73	46	74	46
5	70	46	59	46
6	71	46	55	46

Let us turn now from the development and equating of local tests to their use.

Local Evaluations

The development of the local test was viewed by the research office in this school district as an opportunity to improve the local evaluation.

By using raw scores on the tests to evaluate overall performance and analyzing the subtest areas of results, weaknesses could be more accurately identified.

Equating the local test to a nationally normed test fit well into the scheme of things. The equated percentiles would be reported to the state. There would be no need to report percentiles to the school. The local evaluation would use percent correct scores with the reference point being the grade level or passing score. In this way, school improvement could be emphasized through the increased achievement of specific skills rather than reference to a norm group. With objectives based on the loacl tests, real improvement could be more likely measured.

Few school districts in the country have chosen the option of developing their own tests and equating them to national tests. Cost is a consideration. If you use regular personnel and you have a computer system, the costs would not seem to be prohibitive. It is a feasible project for the medium to large school district with trained evaluators on the staff. It does take time, two to three years from the item writing stage to field testing the final adoption.

APPENDIX

SCHOOL= VICTORY EXTENDED ITEM ANALYSIS

QUESTION NUMBER		RESPONSE COUNT A	B	C	D	E	MR	NR	RESPONSE PERCENT A	B	C	D	E	MR	NR	DIF IND	VAL IND
15	UB7	0	33*	0	1	0	0	0	.0	97.1*	.0	2.9	.0	.0	.0		
	LB7	2	19*	0	0	0	0	4	7.4	70.4*	.0	.0	.0	.0	22.2		
	TOT	2	87*	0	1	0	0	4	2.1	90.6*	.0	1.0	.0	.0	6.3	.91	.85
16	UB7	1	0	0	33*	0	0	0	2.9	.0	.0	97.1*	.0	.0	.0		
	LB7	0	1	1	19*	0	0	4	.0	3.7	3.7	70.4*	.0	.0	22.2		
	TOT	1	1	1	86*	0	0	7	1.0	1.0	1.0	89.6*	.0	.0	7.3	.90	.85
17	UB7	0	0	34*	0	0	0	0	.0	.0	100.0*	.0	.0	.0	.0		
	LB7	0	0	19*	2	0	0	4	.0	.0	70.4*	7.4	.0	.0	22.2		
	TOT	0	2	88*	2	0	0	7	.0	2.1	88.8*	2.1	.0	.0	7.3	.89	.85
18	UB7	1	0	0	32*	0	0	1	2.9	.0	.0	94.1*	.0	.0	2.9		
	LB7	3	0	1	16*	0	0	7	11.1	.0	3.7	59.3*	.0	.0	25.9		
	TOT	5	5	2	78*	0	0	9	5.2	5.2	2.1	78.1*	.0	.0	9.4	.78	.80
19	UB7	0	0	33*	1	0	0	0	.0	.0	97.1*	2.9	.0	.0	.0		
	LB7	2	2	10*	0	0	0	4	33.3	7.4	37.0*	.0	.0	.0	22.2		
	TOT	14	2	70*	1	0	1	6	14.7	2.1	72.9*	1.0	.0	1.0	6.3	.73	.73
20	UB7	0	0	34*	0	0	0	0	.0	.0	100.0*	.0	.0	.0	.0		
	LB7	11	3	5*	1	0	0	7	40.7	11.1	18.5*	3.7	.0	.0	25.9		
	TOT	19	3	68*	2	0	0	7	19.8	3.1	67.7*	2.1	.0	.0	7.3	.68	.82
21	UB7	0	2	1	31*	0	0	0	.0	5.9	2.9	91.2*	.0	.0	.0		
	LB7	0	9	1	10*	0	0	7	.0	33.3	3.7	37.0*	.0	.0	25.9		
	TOT	0	10	2	68*	0	0	7	.0	10.8	2.1	70.8*	.0	.0	7.3	.71	.87
22	UB7	0	32*	1	1	0	0	0	.0	94.1*	2.9	2.9	.0	.0	.0		
	LB7	6	9*	1	2	0	1	5	22.2	33.3*	3.7	7.4	.0	3.7	29.6		
	TOT	12	68*	3	4	0	1	8	12.5	70.8*	3.1	4.2	.0	1.0	8.3	.71	.64
23	UB7	0	0	34*	0	0	0	0	.0	.0	100.0*	.0	.0	.0	.0		
	LB7	4	4	8*	3	0	0	8	14.8	14.8	29.6*	11.1	.0	.0	29.6		
	TOT	7	4	71*	5	0	0	9	7.3	4.2	74.0*	5.2	.0	.0	9.4	.74	.77
24	UB7	0	0	0	34*	0	0	0	.0	.0	.0	100.0*	.0	.0	.0		
	LB7	4	1	3	8*	0	0	11	14.8	3.7	11.1	29.6*	.0	.0	40.7		
	TOT	8	1	4	70*	0	0	13	8.3	1.0	4.2	72.9*	.0	.0	13.5	.73	.77
25	UB7	0	34*	0	0	0	0	0	.0	100.0*	.0	.0	.0	.0	.0		
	LB7	1	14*	0	1	0	0	11	3.7	51.9*	.0	3.7	.0	.0	40.7		
	TOT	1	82*	0	2	0	0	11	1.0	85.4*	.0	2.1	.0	.0	11.5	.85	.69
26	UB7	1	1	0	32*	0	0	0	2.9	2.9	.0	94.1*	.0	.0	.0		
	LB7	0	3	1	11*	0	0	12	.0	11.1	3.7	40.7*	.0	.0	44.4		
	TOT	1	5	1	77*	0	0	12	1.0	5.2	1.0	80.2*	.0	.0	12.5	.80	.61
27	UB7	0	1	32*	1	0	0	0	.0	2.9	94.1*	2.9	.0	.0	.0		
	LB7	1	2	6*	3	0	0	15	3.7	7.4	22.2*	11.1	.0	.0	55.6		
	TOT	1	3	73*	4	0	0	15	1.0	3.1	76.0*	4.2	.0	.0	15.6	.76	.73
28	UB7	34*	0	0	0	0	0	0	100.0*	.0	.0	.0	.0	.0	.0		
	LB7	6*	5	2	2	0	0	13	18.5*	18.5	7.4	7.4	.0	.0	48.1		
	TOT	71*	6	2	3	0	0	14	74.0*	6.3	2.1	3.1	.0	.0	14.6	.74	.88

Appendix